Developing a Controlled Vocabulary for Education as a Health Care Intervention

Malinda Peeples, RN, MS, CDE
Johns Hopkins University School of Medicine
Division of Health Sciences Informatics

NLM Rotation for Medical Informatics Trainee Summer 2004
Mentors: Olivier Bodenreider M.D., Ph.D.
Dina Demner-Fushman, M.D., Ph.D.

Abstract Self-management of chronic disease is a responsibility of the individual who has the disease. Education about the treatments, monitoring, and other activities are key to the successful self-care and effective quality of life. Education about chronic disease occurs in multiple settings, is provided by various healthcare providers, and is delivered by many different methods. This project describes two approaches that used the UMLS and MEDLINE to develop a controlled vocabulary of education as an intervention. A test vocabulary of diabetes education programs was initiated and proof-of- concept was demonstrated.

Introduction

Chronic disease is a significant public health burden in the United States. It affects over 90 million people and accounts for more than 75% of annual medical care costs[1]. As the policy makers and health care providers work to determine best allocation of resources and design delivery models that result in the best health outcomes, clinicians are being challenged to provide evidence of effectiveness and best practice. Chronic disease care occurs in multiple care settings by multiple providers, with multiple treatment strategies. The complexity of chronic disease care requires much care and coordination. Models of chronic disease care are evolving such as the one developed by Dr. Ed Wagner and the Improving Chronic Care Organization. The Chronic Care Model identifies the essential elements of a health care system that encourage high-quality chronic disease care. These elements are the community, the health system, self-management support, delivery system design, decision support and clinical information systems. [2] This project is focused on the self-management support component of this model.

People with chronic disease are primarily responsible for the day to day management of the disease, often requiring complex routines, administration of medications, and monitoring of certain biologic indicators. Learning how to perform these activities and having support for these activities is crucial to impact positives outcomes and help these people to benefit from the best quality of life. Health or patient education is a key intervention that supports the effective self-management and can be defined as imparting knowledge, attitudes and skills with the specific goal of changing behavior, increasing adherence with therapy, to improve health. [3]

Currently, there are many methods and models of patient education – from health information to formal disease specific education programs. Some examples include the Information RX, an initiative of the National Library of Medicine, (NLM) where physicians make referral to information sources on-line at MedlinePlus.gov, an authoritative, Internet site for health information; the National Diabetes Education (NDEP), a partnership of the National Institutes of Health, the Centers for Disease Control and Prevention, and more than 200 public and private organizations providing information to professionals and individuals with diabetes ; disease specific education programs such as asthma, disease management companies, individual providers. Many different health care providers are involved: physicians, nurses, dietitians, pharmacists,

librarians, public health workers, lay health workers, and faith based workers. Also, caregivers and family members are often seeking health information and education to be informed and supportive of the person who has the chronic disease. Therefore, education is an essential health care intervention for chronic disease care. Effective descriptions, measurement, and reporting of the education intervention from the perspective of the provider as well as the person with chronic disease can benefit not only the individuals with chronic disease but also the health care systems, and ultimately the national health care costs.

The purpose of this project is to explore the application of an informatics approach to improving the description of education as a health care intervention. The primary hypothesis is that a structured controlled vocabulary of education as an intervention will support: electronic outcomes reporting; literature indexing (evidence gathering); a common language to support provider and patient communication; and integration of education into the electronic medical record and other entities.

Background

A controlled vocabulary is effective for supporting various information processing activities, including indexing, searching, and aggregating data for classification and evaluation of patient care outcomes[4]. Typically vocabularies are developed with a purpose and scope usually related to a specific domain[5]. Determining a process for developing a vocabulary for a domain that crosses multiple health disciplines, multiple diseases entities, and multiple care processes can be quite challenging.

This project was designed to be proof-of-concept for using existing vocabularies to seed a new vocabulary. The key benefit expected from this effort is a controlled vocabulary that takes advantage of the robustness and richness of existing vocabularies and links the new vocabulary to work already done. However, for terms not contained in existing vocabularies, other methods will be explored. The goal of a vocabulary of education as an intervention is to more explicitly identify and characterize the complexity of the intervention through concepts and relationships.

The objectives of this phase of the study were to pilot a process to develop a test vocabulary from existing vocabularies and to explore a text-based method of defining missing terms from the biomedical literature in the domain of diabetes education using resources available at the National Library of Medicine (NLM).

The Unified Medical Language System (UMLS), a public service of the U.S. National Library of Medicine (NLM), is comprised of knowledge sources and application software tools that support electronic system developers for information processing activities for biomedical and health data. A component of the UMLS, the Metathesaurus, is a very large, multipurpose, and multilingual thesaurus containing concepts, their associated concepts and relationships among them from over 100 source biomedical and health vocabularies [6]

MEDLINE® (Medical Literature, Analysis, and Retrieval System Online) is NLM's bibliographic database that contains over 12 million references to journal articles in life sciences with a concentration on biomedicine. "Citations

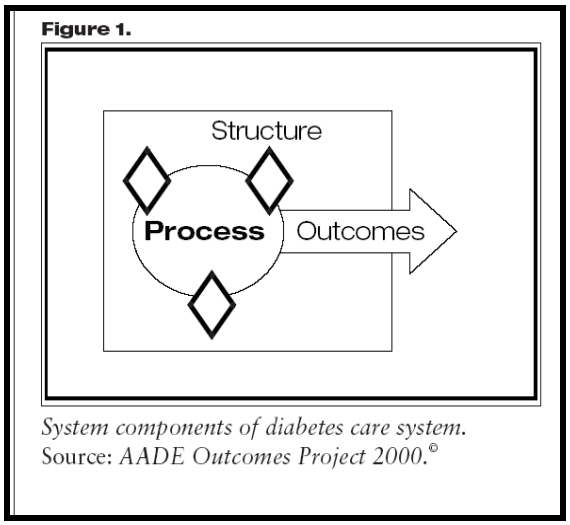
from over 4,600 worldwide journals currently in 30 languages; 40 languages for older journals cited back to 1966. About 52% of current cited articles are published in the U.S.; for the time period 1997-2001, nearly 89% of cited articles are published in English and about 76% have English abstracts written by authors of the articles. Citations for MEDLINE are created by the NLM, international partners, and collaborating organizations." MEDLINE can be searched using NLM's controlled vocabulary MeSH, or by author name, title word, text word, journal name, phrase, or any combination of these. The result of a search is a list of citations (including authors, title, source, and often an abstract) to journal articles.[7]

For the pilot process, a narrow project purpose and scope was defined for the vocabulary. The author is involved in a national project to develop an outcomes repository and reporting service for diabetes education programs in collaboration with the American Association of Diabetes Educators.

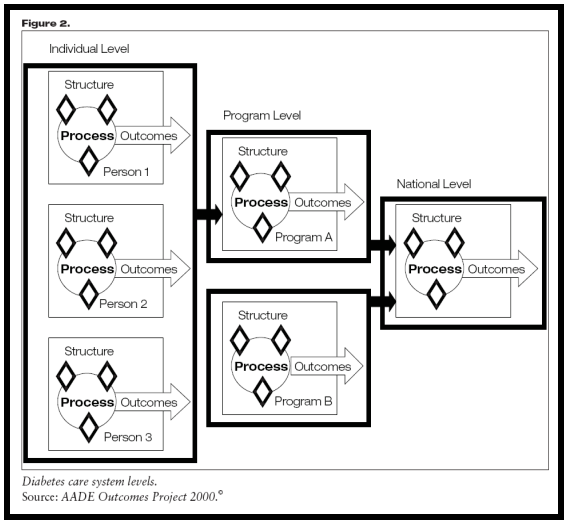
The conceptual framework of the National Diabetes Education Outcomes System (NDEOS) is significant to this vocabulary project. The framework was based on systems theory and the work of Donebedian in the area of quality management. The structure, process, and outcomes framework addresses how components of a system are organized (structure) to perform specific functions (processes) to meet certain objectives or results (outcomes). NDEOS integrates the system of diabetes self-management education into the overall diabetes care system.

The National Diabetes Education Outcomes System (NDEOS) is a complete system of standardized measures, measurement tools, and reporting for individual, program, and national level outcomes of diabetes education. The systems architecture of NDEOS includes data acquisition methods of web-based, optical scan, and telephonic technologies uploaded into an industrial strength database repository. The vision of the system, which is currently being implemented, is to acquire longitudinal data on diabetes programs and ultimately benchmark program characteristics, delivery methods, and outcomes of education and care.

Another significant feature of the NDEOS is the measurement and reporting at the individual, program, and national levels. At the individual level, the person with diabetes, the clinicians, and their activities are the focus of definition and measurement. At the program level, the organizational infrastructure and the management that guide the diabetes education are the focal points. The program can range from an individual educator to a multi-site diabetes center with an interdisciplinary team and includes all the participants served by the program. The national level is an aggregate of all the participating programs or centers and can also include professional organizations representing the providers of the interventions[8]. The following figures integrate the two concepts of structure-process-outcomes of the diabetes care system (figure 1),



and individual, program and national levels of the system (figure 2).



Methodology

A test vocabulary to support education program outcomes reporting will be the result of the pilot process. Focusing the vocabulary to a defined scope will facilitate moving the iterative process through the various steps quickly and help to determine the full process. The use case for the test vocabulary will be Diabetes Self-Management Education Program (DSME) from the perspective of the program manager. This focused purpose and scope will limit the project to diabetes, a representative chronic disease affecting all age groups and involving multiple providers in the education and care of the disease.

Two approaches for use of the NLM resources evolved as efficient and effective ways of building the core vocabulary for education as a health intervention. These two methods can be characterized as the "top-down" and the "bottom-up" approaches. First, the "top-down" method refers to exploring the UMLS for existing concepts and their relationships. Next, the "bottom-up" method refers to text-based searching of the biomedical literature using Medline for concepts not represented in the UMLS. For the remainder of the paper, these methods will be referred to as the UMLS Method.

Since this process is evolving and iterative, the methods and the results of each step will be presented together and sequentially for ease in understanding. Recognize that some of the steps as explicitly represented may have evolved from several iterations but are presented as one for ease of understanding for the reader.

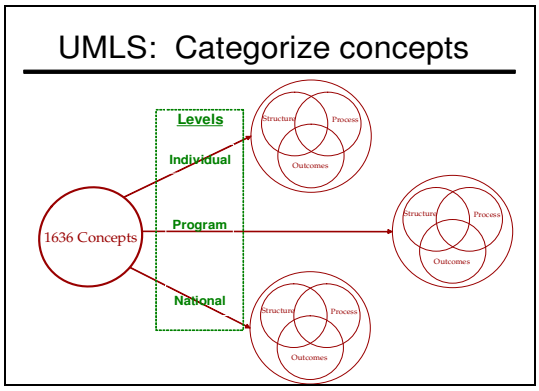
UMLS Method and Results

1. Search the UMLS for concepts – Initially 26 candidate concepts related to diabetes education programs were identified by a domain expert (the author) using the perspective of the program manager. When mapped to the UMLS, 16 seed concepts were identified with 2038 related concepts found. These relationships were hierarchical (ancestors on two generations, direct descendants) or associative relation ("other related

concepts", co-occurring MeSH descriptors in MEDLINE)

2. Manually review the concepts - The 2038 concepts were manually reviewed for relatedness to the domain of education programs. Initially 402 concepts were identified as non-relevant to the vocabulary scope and purpose. "Interesting" terms, or terms that might have some relationship to education at the individual or national levels were archived for later review. Semantic types were used in the review for clarification of the use of the terms in the source vocabularies. However, the meaning of the concept in the education vocabulary may not be of the same semantic type, and concepts were not excluded based on semantic type – this was simply noted. Not surprising, the most common semantic types were *Health Care Activity* and *Educational Activity*.

3. Categorize the concepts – The 1636 remaining concepts were categorize using the framework of structure, process, and outcomes previously described. As the concepts were categorized into this framework, they were filtered through the levels into individual, program, and national. Again the focus was at the program level, with the other two levels being categorized secondarily. Within each level, there were concepts that could be placed into more than one category.

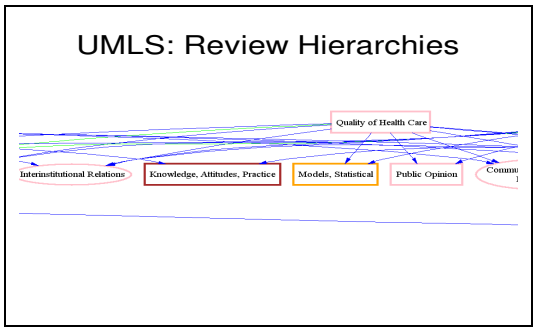


3. Sort the concepts – For each level, the categories were reviewed and sorted into non overlapping sets. To expedite this process, the levels and components were divided into subcategories so that for each level, the structure, process, and outcome components had three subcategories. At the completion of this step, the concepts for the program level numbered 268 – a reasonable number for moving to the next step.

UMLS: Sort Concepts

Levels	Structure	Process	Outcomes
Individual	0	207	152
Program	76	122	70
National	11	27	62

4. Investigate the hierarchical relationships of the categorized concepts - For this step, the outcomes category of the program level was selected for review. Direct and indirect relationships were identified. Direct relationships imply direct inheritance or association whereas the indirect relationship indicated missing concepts between the linked concepts. These relationships will need further investigation to explore for additional concepts that should be added to the vocabulary or point up missing concepts that are needed to establish a different relationship for the education vocabulary.

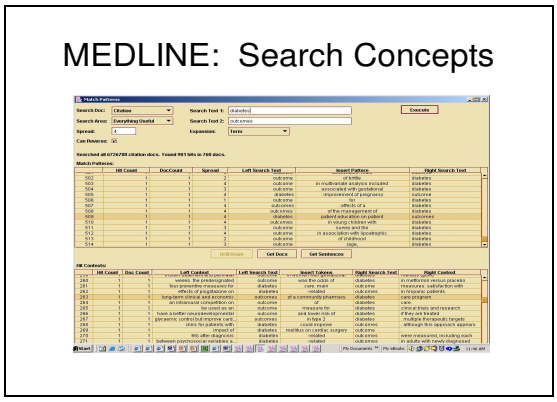


MEDLINE Method: A Text-Based Approach

One option for increasing the vocabulary to include as many concepts as appropriate is to search the biomedical literature for terms that are not included in the source vocabularies of the UMLS. MEDLINE contains over citations from over 4,600 journals. Given the rich source of text, reviewing key abstracts related to chronic disease education may yield additional concepts that will enrich the vocabulary. The goal
of this approach is to determine the most efficient and effective process for reviewing text with the result being a methodology to identify missing concepts especially at the "bottom" level or leaf level in the hierarchy.

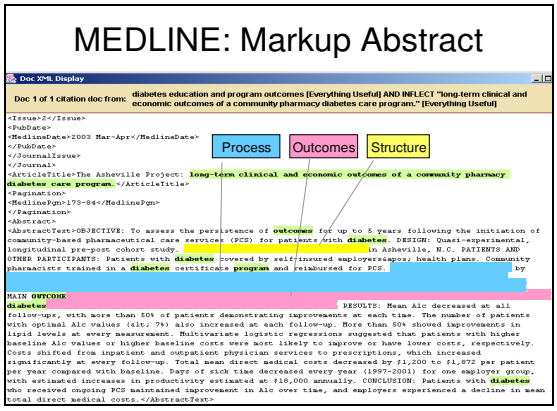
For the Medline method, we used the NLM developed search engine, SEE, Search Explorer. This powerful search engine utilizes some components of natural language processing and these components were applied to this project in the following way.

Step 1. Key concepts were searched and strings of text were identified. Of interest were texts before and after the concepts.



Step 2. Abstracts containing these key strings of text were obtained.

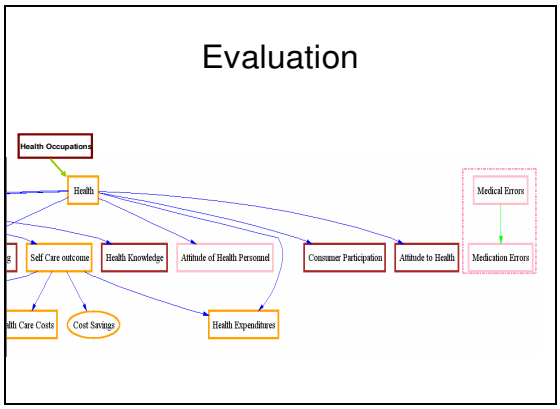
Step 3. Abstracts were marked up using the structure, process, and outcomes framework. Time limited the completion of this step of the project, but the next steps will include review of the marked up text for concepts that are missing. For this step to be the most meaningful, other domain experts will need to be involved in the markup for inter-rater reliability.



Evaluation

Evaluation of the vocabulary should address the completeness, consistency, and quality criteria. Completeness and consistency refer to missing concepts and missing relations. Using the hierarchies from Step 4 of the UMLS method, both criteria can be evaluated. Lack of hierarchical consistency is identified by

isolated concepts that have no identified relationship to the other concepts. In the following example, medical errors is directly related to medication errors but these concepts have neither direct, indirect, or associated relationships to any other concepts. Missing concepts are apparent in the concepts that are indirectly linked. For instance, as noted below, there is at least one, and maybe more, missing concepts between health and attitude to health. Further investigation of these two criteria will inform the vocabulary development – the concepts identified through this process may or may not have relevance to the vocabulary purpose and scope.



A second evaluation criterion is that of quality. Semantic properties may inform the vocabulary development. Agreement with the concept use and the semantic type identified in the UMLS may or may not be significant. However, at this stage of the process it will be noted. Outlier semantic types, identified by the low frequency of occurrence should be evaluated. An example of this was noted in the Program Level, Outcome Component: The concept, C0001811 Aging, had the semantic types *Organism Function* and *Temporal Concept.* With a frequency of 1, does aging belong in the Outcome Component? Perhaps it remains as a concept with a different semantic type or moves to another component category to represent a population and this is yet to be determined.

Conclusion

As proof- of- concept, the project was successful. The resulting methodology, while incomplete, appears to provide a structured approach to building a vocabulary that utilizes the robust UMLS and the rich Medline text. Both approaches are necessary to complete a vocabulary; however a limitation of this approach assumes that the concepts needed exist in the UMLS and biomedical literature.

The pilot project yielded a partial test vocabulary for education programs. This test vocabulary will provide the foundation for continued work on a controlled vocabulary for education as a health intervention. However, the test vocabulary must be completed and evaluated in a formal process.

Continued work on the UMLS method will refine the concepts and their relationships for the test vocabulary.

Exploration of the use of text processing tools such as MetaMap and other automation of concept extraction will be explored with the biomedical literature and perhaps clinical text or program certification data.

The involvement of stakeholder groups in both content development and evaluation will be explored. Consideration will be given to discussion with chronic disease groups other than diabetes, perhaps asthma.

Once developed, the controlled vocabulary can be integrated into the NDEOS system, given that an appropriate infrastructure and collaboration with key stakeholder organizations exist to support the ongoing maintenance of such vocabulary.

References

[1] Chronic Disease Program, Atlanta, GA. Center for Disease Control. Available at http://www.cdc.gov/nccdphp/overview.htm

[2] Chronic Care Model, Seattle, WA. Available http://www.improvingchroniccare.org/change/model/components.html#citation

[3] Consumer and Patient Health Information Section of the Medical Library Association, 1996 NEED web

[4] Lorence DP, Spink A. Semantics and the medical web: a review of barriers and breakthroughs in effective healthcare query. Health Info Libr J. 2004 Jun; 21(2): 109-16.

[5] Elkin PL, Brown SH, Carter J, et al. Guideline and quality indicators for development, purchase and use of controlled health vocabularies. International Journal of Medical Informatics 2202; 68: 175-186.

[6]Unified Medical Language System (UMLS), Bethesda, MD. National Library of Medicine. Available from URL: http://www.nlm.nih.gov/research/umls/about_umls.html.

[7] Medline, Bethesda, MD. National Library of Medicine. Available from URL: http://www.nlm.nih.gov/pubs/factsheets/medline.html.

[8] Peeples MM, Mulcahy KM, Tomky D. The Conceptual Framework for the NDEOS. . The Diab Educ. 2001; 27(4): 547-552.